

# Simultaneous Local Binary Feature Learning and Encoding for Homogeneous and Heterogeneous Face Recognition

Jiwen Lu , Senior Member, IEEE, Venice Erin Liong, Student Member, IEEE, and Jie Zhou, Senior Member, IEEE

**Abstract**—In this paper, we propose a simultaneous local binary feature learning and encoding (SLBFLE) approach for both homogeneous and heterogeneous face recognition. Unlike existing hand-crafted face descriptors such as local binary pattern (LBP) and Gabor features which usually require strong prior knowledge, our SLBFLE is an unsupervised feature learning approach which automatically learns face representation from raw pixels. Unlike existing binary face descriptors such as the LBP, discriminant face descriptor (DFD), and compact binary face descriptor (CBFD) which use a two-stage feature extraction procedure, our SLBFLE jointly learns binary codes and the codebook for local face patches so that discriminative information from raw pixels from face images of different identities can be obtained by using a one-stage feature learning and encoding procedure. Moreover, we propose a coupled simultaneous local binary feature learning and encoding (C-SLBFLE) method to make the proposed approach suitable for heterogeneous face matching. Unlike most existing coupled feature learning methods which learn a pair of transformation matrices for each modality, we exploit both the common and specific information from heterogeneous face samples to characterize their underlying correlations. Experimental results on six widely used face datasets including the LFW, YouTube Face (YTF), FERET, PaSC, CASIA VIS-NIR 2.0, and Multi-PIE datasets are presented to demonstrate the effectiveness of the proposed methods.

**Index Terms**—Face recognition, heterogeneous face matching, feature learning, binary feature, compact feature, biometrics

## 1 INTRODUCTION

FACE recognition is a classical and representative computer vision problem, and a variety of face recognition algorithms have been proposed in the literature [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Generally, there are two important procedures for a practical face recognition system: face representation and face matching. Face representation aims to extract discriminative feature descriptors to make face images/videos more separable, and face matching is to design effective classifiers/models to differentiate different face samples. Practical face recognition systems are usually affected by many variations such as occlusions, poses, illuminations, expressions and resolutions, which usually cause large intra-class variations. Hence, how to obtain robust face representations which are invariant and robust to many real-world variations is key challenge in face recognition systems. Existing face representation methods can be

mainly classified into two categories: holistic feature representation [2], [3] and local feature representation [5], [12]. Representative holistic feature representation methods include principal component analysis (PCA) [2] and linear discriminant analysis (LDA) [3], and typical local feature descriptors are local binary pattern (LBP) [5] and Gabor features [12]. While many face descriptors have been proposed in the literature [6], [7], [8], [9], [10], [11], most of them are hand-crafted and usually require strong prior knowledge to design. Moreover, some of them are computationally expensive, which may limit their practical applications.

Feature learning has been successfully applied for face recognition. For example, Cao et al. [13] presented a learning-based (LE) feature representation method by applying the bag-of-words (BoW) framework. Hussain et al. [14] proposed a local quantized pattern (LQP) and Lei et al. [15] proposed a discriminant face descriptor (DFD) method to learn LBP-like features. Sun et al. [16] proposed a deep convolutional neural networks method to learn face representations. However, most of them learn real-valued face feature descriptors. For face recognition, binary features are more robust to local changes in face images because small variations caused by varying expressions and illuminations can be eliminated by quantized binary codes.

In this paper, we propose a simultaneous local binary feature learning and encoding (SLBFLE) method for face recognition. Fig. 1 illustrates the basic idea of our proposed approach. Motivated by the fact that local binary features are robust to local changes such as varying illuminations

- J. Lu and J. Zhou are with the Department of Automation, Tsinghua National Laboratory for Information Science and Technology (TNList), State Key Lab of Intelligent Technologies and Systems, Tsinghua University, Beijing 100084, China. E-mail: {lujiwen, jzhou}@tsinghua.edu.cn.
- V.E. Liong is with the Interdisciplinary Graduate School, Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore 639798, Singapore. E-mail: veniceer001@e.ntu.edu.sg.

Manuscript received 23 Nov. 2016; revised 6 July 2017; accepted 4 Aug. 2017.  
Date of publication 8 Aug. 2017; date of current version 11 July 2018.  
(Corresponding author: Jiwen Lu.)

Recommended for acceptance by M. Tistarelli.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TPAMI.2017.2737538

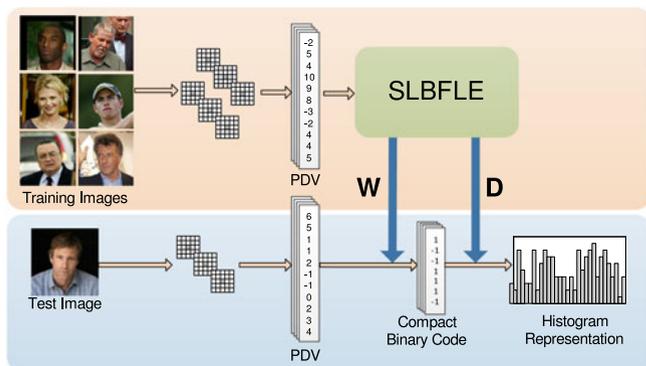


Fig. 1. The basic idea of the proposed SLBFLE approach for face representation. For each training face image, we extract pixel difference vectors (PDVs) and jointly learn a discriminative mapping  $W$  and a dictionary  $D$  for feature extraction. The mapping is to project each PDV into a low-dimensional binary vector, and the dictionary is used as the codebook for feature local encoding. For each test image, the PDVs are first extracted and encoded into binary codes using the learned feature mapping, and then converted as a histogram feature with the learned dictionary.

and expressions [5], [17], we aim to learn compact binary codes directly from raw pixels for face representation. Unlike previous binary feature descriptors such as LBP and discriminant face descriptor [15] which use a two-stage feature extraction approach, our proposed SLBFLE jointly learns binary codes and the codebook for feature encoding over local face patches so that discriminative information from raw pixels in face images of different identities can be jointly learned with a one-stage procedure. These local face patches used for training are extracted from different regions of the face image to ensure all variations of local patches are utilized. For each local patch, the pixel difference vectors (PDVs) are extracted to provide representative information of edges and line patterns which are suitable for simultaneous binary code and codebook learning. The key motivation of our simultaneous learning strategy is that some useful information for feature encoding may be compromised in the feature learning stage if they are employed sequentially because their key objectives are different. Specifically, feature learning can be considered as a Hamming metric learning problem and feature encoding is a dictionary learning problem. Hence sequential learning only provides an one-way interaction where binary codes are learned independently first and then used to learn the codebooks for feature encoding. Therefore, these learned binary codes cannot receive the feedback in the codebook learning stage. In our simultaneous learning framework, the binary code learning procedure also considers how to generate efficient codebooks to have representative face representation. By doing so, more robust and discriminative feature descriptors can be extracted. In addition, we propose a coupled simultaneous local binary feature learning and encoding (C-SLBFLE) for heterogeneous face recognition. Unlike existing other coupled feature learning techniques which learn different specific structures for each modality, our C-SLBFLE learns compact representations which compose of a specific latent representation obtained from each modality, and a common latent representation shared across modalities, respectively. By doing so, we obtain a representation which aims to maximize the similar and consistent information among these two modalities as well exploit specific complementary information for each modality.

The contributions of this work are summarized as follows:

- 1) We propose an unsupervised feature learning method that jointly perform compact feature learning and encoding to obtain robust and discriminative face representations.
- 2) We propose a coupled feature learning method with shared structural learning to exploit both the common and specific information from heterogeneous data for heterogeneous face matching.
- 3) We conduct extensive face recognition experiments on six widely used face datasets including LFW, YouTube Face (YTF), FERET, PaSC, CASIA NIR-VIS, and Multi-PIE where our results clearly demonstrate the effectiveness of the proposed methods.

## 2 RELATED WORK

In this section, we briefly review three related topics: 1) feature learning for homogeneous and heterogeneous face recognition, and 2) binary code learning.

### 2.1 Feature Learning

*Homogeneous Face Recognition.* Representative feature learning methods include sparse auto-encoders [18], convolutional neural networks [19], independent subspace analysis [20], and reconstruction independent component analysis [21]. Feature learning has also been successfully applied for face recognition. For example, Cao et al. [13] presented a learning-based feature representation method by applying the bag-of-words (BoW) framework. Hussain et al. [14] proposed a local quantized pattern method by modifying the LBP method with a learned coding strategy. Lei et al. [15] proposed a discriminant face descriptor method by learning an image filter using the LDA criterion to obtain LBP-like features. Sun et al. [16] proposed a deep convolutional neural networks method to learn face representations. Recently, Taigman et al. [22] introduced a DeepFace method which learn supervised face representation with 4,000,000 labeled face samples by using the deep convolutional neural networks. Parkhi et al. [23] presented A deep convolutional neural networks method with the triplet loss function, where 2.6M labeled face images were used to train the deep model. However, most of these methods learn real-valued face feature descriptors. For face recognition, binary features are more robust to local changes in face images because small variations caused by varying expressions and illuminations can be eliminated by quantized binary codes.

*Heterogeneous Face Recognition.* Heterogeneous face recognition (HFR) aims to match face samples captured from various devices or environments. A representable example is the photo-sketch face matching which is useful for law enforcement, where a sketch of a suspect is given as probe and RGB face images are provided as gallery for matching. Another example is the NIR-VIS matching where we try to match face images captured from RGB cameras (VIS) and near-infrared imaging devices (NIR) from different environments. While some feature learning methods have been proposed for homogeneous face recognition, these methods are still not robust when face images are captured across different

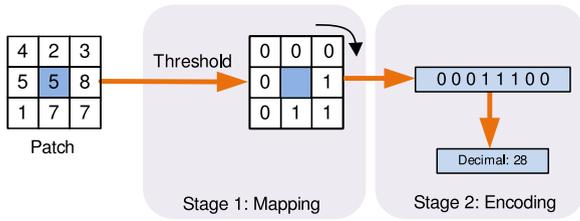


Fig. 2. The basic idea of the LBP method, where a two-stage procedure is used for local feature extraction: feature mapping and feature encoding. For the feature mapping stage, the difference between the central pixel and the neighboring pixels are computed and binarized with a fixed threshold. For the feature encoding stage, the mapped binary codes are encoded as a real value by using a hand-crafted pattern coding strategy.

devices or environments because large appearance gap usually occurs. How to extract common properties and reduce this gap is the key challenge in heterogenous face recognition. Recently, a variety of feature learning methods [24], [25], [26] have also been proposed for heterogenous face recognition. For example, Jin et al. [26] learned representative features by training a coupled filters which maximizes the inter-class variations and minimize the intra-class variations. Saxena and Verbeek [24] used a CNN model with a shared layer learned from a soft-max criterion to obtain common features. Yi et al. [25] extracted Gabor features from face landmarks and performed shared representation learning to reduce the modality gap. Differently, in this work, we learn specific and common latent spaces to obtain similar information and exploit specific complimentary information, respectively.

### 2.2 Binary Code Learning

A variety of binary code learning methods have been proposed in recent years [27], [28], [29]. For example, Weiss et al. [29] proposed a binary coding learning approach for image search. Norouzi et al. [30] improved it by using a triplet ranking loss optimization criterion. However, most existing binary code learning methods are developed for scalable similarity search [28]. While binary features such as LBP and Haar-like features have been used in face recognition, most of them are hand-crafted. There have been some recent work which employs binary code learning for face representation and recognition [31], [32], [33]. For example, Zhang et al. [32] and Rastegari et al. [33] learned binary codes based on variants of the fisher criterion. However, these binary codes are learned holistically and not in feature level. More recently, Lu et al.[31] introduced a compact binary feature descriptor (CBFD) which learned binary face descriptors at the feature level. However, CBFD performed feature and codebook learning separately, so that some useful information for codebook learning may be compromised in the binarization stage.

## 3 PROPOSED APPROACH

In this section, we first review the LBP method and present the proposed SLBFLE method. Then we show how to use SLBFLE for face representation. Lastly, we present the proposed C-SLBFLE method for heterogenous face recognition.

### 3.1 Review of LBP

LBP is an effective feature descriptor in face recognition [5]. For each pixel in face image, LBP first computes the

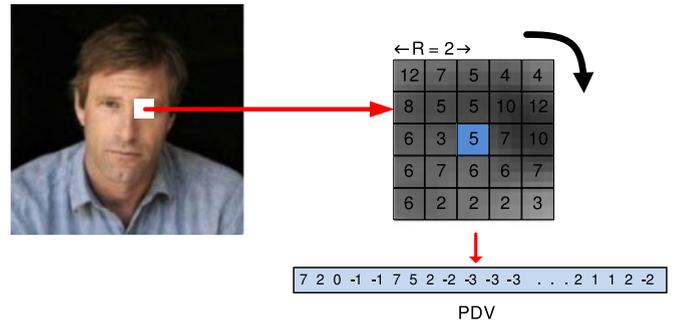


Fig. 3. An illustration to show how to extract pixel difference vectors (PDV) from the original face image. Given a face patch whose size is  $(2R + 1) \times (2R + 1)$ , we first compute the difference between the central pixel and the neighboring pixels. Then, these differences are considered as a PDV. In this figure,  $R$  is selected as 2, so that there are 24 neighboring pixels selected and the PDV is a 24-dimensional feature vector.

difference between the central pixel and the neighboring pixels and binarizes the difference with a fixed threshold. Second, these binary bins are encoded as a real value by using a hand-crafted pattern coding strategy. Fig. 2 illustrates the basic idea of LBP, where two individual stages are used for feature representation.

There are two shortcomings in LBP: 1) both the binarization and feature encoding stages are hand-crafted, which are not optimal for local feature representation; 2) a two-stage procedure is used in LBP, which is not effective enough because some useful information for codebook learning may be compromised in the binarization stage. To address this, we propose a SLBFLE method to learn a discriminative mapping and a compact codebook for feature mapping and encoding jointly, so that more data-adaptive information can be exploited in the learned features. The following describes the details of the proposed method.

### 3.2 SLBFLE

As aforementioned, our SLBFLE aims to jointly learn a feature mapping and a dictionary for feature mapping and encoding. While our SLBFLE method is unsupervised, it still has strong discriminative power because raw pixels are extracted from face images of different identities which contribute to learning a discriminative feature mapping. Moreover, the learned binary codes can well describe how pixel values change over local patches and implicitly encode important visual patterns such as edges and lines in face images. Also, the learned dictionary can well encode the learned binary codes so that some noisy information can be well alleviated.

Let  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{d \times N}$  be a set of  $N$  face image samples, where  $\mathbf{x}_n \in \mathbb{R}^d$  ( $1 \leq n \leq N$ ) is a pixel difference vector extracted from an original face image. Fig. 3 illustrates how to extract a PDV for a given face patch. Compared with the original raw pixel patch, PDV measures the difference between the central pixel and the neighboring pixels within a patch, so that it can better describe how pixel values change spatially and implicitly encode important visual patterns such as edges and lines in face images. Assume there are  $K$  hash functions to be learned in SLBFLE, which map and quantize each  $\mathbf{x}_n$  into a binary vector  $\mathbf{b}_n = [b_{n1}, \dots, b_{nK}] \in \{-1, 1\}^{K \times 1}$ , so that the binary codes

are learned automatically rather than using an empirical thresholding method. Let  $\mathbf{w}_k \in \mathbb{R}^d$  be the projection vector for the  $k$ th function, the  $k$ th binary code  $b_{nk}$  of  $\mathbf{x}_n$  can be computed as follows:

$$b_{nk} = \text{sgn}(\mathbf{w}_k^\top \mathbf{x}_n), \quad (1)$$

where  $\text{sgn}(v)$  equals to 1 if  $v \geq 0$  and  $-1$  otherwise.<sup>1</sup>

Having obtained binary codes for these PDVs in the training set, we also require a codebook to pool those binary codes in each face image into a histogram feature. Previous methods applied the  $K$ -means algorithm to learn the codebook [13], [14], [15]. However, some useful information for codebook learning may be compromised in the mapping learning stage if they are learned sequentially. In this work, we learn them simultaneously so that discriminative information can be jointly exploited.

Let  $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_C]$  and  $\mathbf{A} = [\alpha_1, \alpha_2, \dots, \alpha_N]$  be the dictionary and the corresponding representation coefficient matrix, respectively, where  $\mathbf{d}_c \in \mathbb{R}^{K \times 1}$  ( $1 \leq c \leq C$ ) is the  $c$ th atom in the dictionary,  $C$  is the total number of atoms in the dictionary,  $\alpha_n \in \mathbb{R}^{C \times 1}$  is the representation coefficient for  $\mathbf{x}_n$ . We formulate the following optimization problem

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{D}, \alpha} J &= J_1 + \lambda_1 J_2 \\ &= \sum_{n=1}^N \left( \|\mathbf{b}_n - \mathbf{D}\alpha_n\|_2^2 + \gamma \|\alpha_n\|_1 \right) \\ &\quad + \lambda_1 \sum_{n=1}^N \sum_{k=1}^K \|\mathbf{b}_{nk} - \mathbf{w}_k^\top \mathbf{x}_n\|^2 \\ \text{subject to} \quad &\left\| \sum_{n=1}^N \mathbf{b}_{nk} \right\|^2 = 0, \quad \forall k \\ &\mathbf{b}_n \mathbf{b}_n^\top = \mathbf{I}^{k \times k}, \quad \forall n, \end{aligned} \quad (2)$$

where  $\mathbf{b}_n$  is the binary code vector for  $\mathbf{x}_n$ , and  $b_{nk}$  is the  $k$ th bit of  $\mathbf{b}_n$ ,  $\lambda_1$  is the parameter to balance the importance of different terms.

The objective of  $J_1$  is to learn a dictionary  $\mathbf{D}$  over the binary codes where each binary vector can be sparsely reconstructed by  $\mathbf{A}$ . The goal of  $J_2$  is to minimize the quantization loss between the original real-valued features and the binarized codes, so that most energy of the real-valued PDVs can be preserved in the learned binary codes. The first constraint is to ensure that each feature bit in the learned binary codes is evenly distributed over all the training samples (almost half of them are 1, and the other half are 0), so that the information conveyed by each bit is as large as possible. While the second constraint ensures that each projection vector results to independent and uncorrelated binary vectors.

Let  $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_K] \in \mathbb{R}^{d \times K}$  be the projection matrix. We can map each PDV sample  $\mathbf{x}_n$  into a binary vector as follows:

$$\mathbf{b}_n = \text{sgn}(\mathbf{W}^\top \mathbf{x}_n). \quad (3)$$

Then, (2) can be re-written into the matrix form as follows:

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{D}, \mathbf{A}} J &= J_1 + \lambda_1 J_2 \\ &= \|\mathbf{B} - \mathbf{D}\mathbf{A}\|_F^2 + \gamma \|\mathbf{A}\|_1 \\ &\quad + \lambda_1 \|\mathbf{B} - \mathbf{W}^\top \mathbf{X}\|_F^2 \\ \text{subject to} \quad &\mathbf{B} \times \mathbf{1}^{N \times 1} = 0 \\ &\mathbf{B}\mathbf{B}^\top = \mathbf{N}\mathbf{I}^{k \times k}, \end{aligned} \quad (4)$$

where  $\mathbf{B} = \text{sgn}(\mathbf{W}^\top \mathbf{X}) \in \{-1, 1\}^{K \times N}$  is the binary code matrix of all the training samples.

While the objective function in (4) is not convex for  $\mathbf{D}$ ,  $\mathbf{A}$ , and  $\mathbf{W}$ , simultaneously, it is convex to one of them when the other two are fixed. We iteratively optimize  $\mathbf{W}$ ,  $\mathbf{D}$  and  $\mathbf{A}$  by using the following iterative approach. We first initialize  $\mathbf{W}$ ,  $\mathbf{D}$  and  $\mathbf{A}$  appropriate parameters and then iteratively update them sequentially as follows:

*Step 1: Learning  $\mathbf{A}$  with fixed  $\mathbf{W}$  and  $\mathbf{D}$ :* when  $\mathbf{W}$  and  $\mathbf{D}$  are fixed, the objective function in (4) can be re-written as follows:

$$\min_{\mathbf{A}} J = \|\mathbf{B} - \mathbf{D}\mathbf{A}\|_F^2 + \gamma \|\mathbf{A}\|_1, \quad (5)$$

Since (5) is non-differentiable due to the sparsity function, standard unconstrained optimization techniques are infeasible and gradient-based methods cannot be applied directly. Instead, we optimize the objective function by decomposing it into a series of individual  $\ell_1$ -regularized least square problem for  $\alpha_n$  as follows:

$$\min_{\alpha_n} J = \sum_{n=1}^N \left( \|\mathbf{b}_n - \mathbf{D}\alpha_n\|_2^2 + \gamma \sum_{j=1}^K |\alpha_n^{(j)}| \right), \quad (6)$$

where  $\alpha_n$  is the  $n$ th column of  $\mathbf{A}$ , and  $|\alpha_n^{(j)}|$  is the  $j$ th element of  $\alpha_n$ . This optimization problem actually reflects a sparse coding problem which can already be solved by several optimization solutions [34], [35]. In this paper, we use the feature sign search algorithm in [34] to optimize  $\alpha_n$  sequentially.

*Step 2: Learning  $\mathbf{D}$  with fixed  $\mathbf{W}$  and  $\mathbf{A}$ :* when  $\mathbf{W}$  and  $\mathbf{A}$  are fixed, the optimization function in (4) can be re-written as the following objective function:

$$\begin{aligned} \min_{\mathbf{D}} J &= \|\mathbf{B} - \mathbf{D}\mathbf{A}\|_F^2 \\ \text{s.t.} \quad &\|\mathbf{d}_c\|_2^2 \leq 1, 1 \leq c \leq C. \end{aligned} \quad (7)$$

The optimization objective function in (7) is a standard  $\ell_2$ -constrained optimization problem. We use the conventional conjugate gradient decent method in [21] to optimize  $\mathbf{D}$ .

*Step 3: Learning  $\mathbf{W}$  with fixed  $\mathbf{D}$  and  $\mathbf{A}$ :* when  $\mathbf{D}$  and  $\mathbf{A}$  are fixed, (4) can be re-written as follows:

$$\begin{aligned} \min_{\mathbf{W}} J &= \|\mathbf{B} - \mathbf{D}\mathbf{A}\|_F^2 \\ &\quad + \lambda_1 \|\mathbf{B} - \mathbf{W}^\top \mathbf{X}\|_F^2 \\ \text{subject to} \quad &\mathbf{B} \times \mathbf{1}^{N \times 1} = 0 \\ &\mathbf{B}\mathbf{B}^\top = \mathbf{N}\mathbf{I}^{k \times k}. \end{aligned} \quad (8)$$

To further simplify the computation and make the formulation tractable, we perform relaxation on several terms. First, the balancing constraint can be relaxed by

1. During learning, the binary codes are set to  $\{-1, 1\}$  instead of  $\{0, 1\}$  to ensure proper centering.

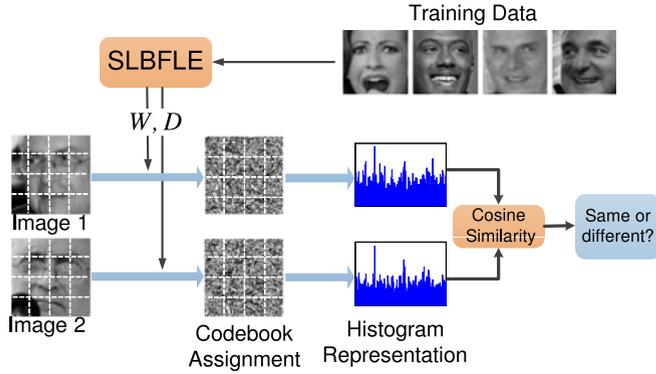


Fig. 4. The flow-chart of the SLBFLE-based face representation approach. For each training face, we first divide it into several non-overlapped regions and learn the feature mapping  $\mathbf{W}$  and dictionary  $\mathbf{D}$  for each region. Then, we applied the learned filter and dictionary to extract histogram feature for each block and concatenated into a longer feature vector for face representation. Finally, the cosine similarity measure is used to measure face similarity for verification.

maximizing the variance as justified in [27]. Second, the binary constraint caused by non-linear  $\text{sgn}(\cdot)$  function is also relaxed as a signed-magnitude projection as justified by [27], [28]. The first constraint can then be simplified as maximizing the signed-magnitude variance of the projection as follows:

$$\sum_k^K \mathbb{E}[\|\mathbf{w}_k^\top \mathbf{x}\|^2] = \text{tr}(\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W}). \quad (9)$$

The second constraint is also relaxed as:

$$\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W} = \mathbf{N} \mathbf{I}^{k \times k}. \quad (10)$$

It can be noticed that the first and second constraint can be combined into one penalty parameterized by  $\lambda_2$ . The overall function can now be rewritten as follows:

$$\begin{aligned} \min_{\mathbf{W}} J &= \|\mathbf{W}^\top \mathbf{X} - \mathbf{D} \mathbf{A}\|_F^2 + \lambda_1 \|\mathbf{B} - \mathbf{W}^\top \mathbf{X}\|_F^2 \\ &\quad - \lambda_2 \text{tr}(\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W}) \\ &= (1 + \lambda_1 - \lambda_2) \text{tr}(\mathbf{W}^\top \mathbf{X} \mathbf{X}^\top \mathbf{W}) \\ &\quad - 2\lambda_1 \text{tr}(\mathbf{B}^\top \mathbf{W}^\top \mathbf{X}) \\ &\quad + 2\text{tr}((\mathbf{W}^\top \mathbf{X})^\top \mathbf{D} \mathbf{A}). \end{aligned} \quad (11)$$

We use the gradient descent method with the curvilinear search algorithm in [36] to solve  $\mathbf{W}$ .

We repeat the above three steps until the algorithm is convergent. Algorithm 1 summarizes the detailed procedure of the proposed SLBFLE method.

### 3.3 SLBFLE-Based Face Representation

Having obtained the feature mapping  $\mathbf{W}$  and the dictionary  $\mathbf{D}$ , we first project each PDV into a low-dimensional binary vector and encode it as a real value. Then, all PDVs within the same face region is represented as a histogram feature. Finally, these features from all blocks within a face are concatenated as the feature representation of the whole face image. Fig. 4 illustrates how to use the proposed SLBFLE for face representation.

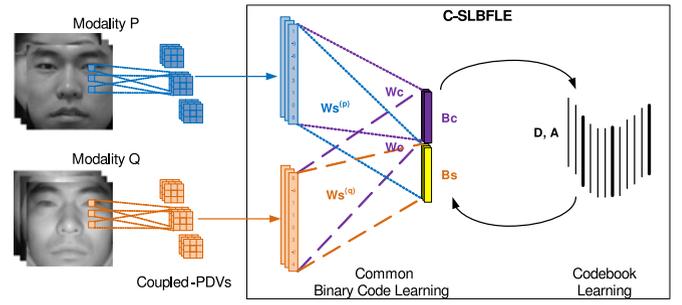


Fig. 5. The flow-chart of the proposed C-SLBFLE method. We first extract the coupled-PDVs from a pair of aligned face images captured from different modalities. These coupled-PDVs are then used to jointly learn the feature mapping  $\mathbf{W}_p$  and  $\mathbf{W}_q$  for the modality  $P$  and  $Q$ , respectively, and codebook  $\mathbf{D}$ . Each feature mapping matrix composes of a common projection matrix,  $\mathbf{W}_c$ , and modal-specific projection matrices  $\mathbf{W}_s^{(p)}$  and  $\mathbf{W}_s^{(q)}$ . The final codebook and projection matrices are then used to extract the face representations during testing stage.

### Algorithm 1. SLBFLE

**Input:** Training set  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ , iteration number  $iter$ , parameters  $\lambda_1$  and  $\lambda_2$ , and binary code length  $K$ .

**Output:** Projection  $\mathbf{W}$ , dictionary  $\mathbf{D}$ , and coefficient matrix  $\mathbf{A}$ .

Step 1 (**Initialization**):

1.1 Initialize  $\mathbf{W}$  as the top  $K$  eigenvectors of  $\mathbf{X} \mathbf{X}^\top$  corresponding to the  $K$  largest eigenvalues.

1.2 Initialize  $\mathbf{D}$  and  $\mathbf{A}$  with arbitrary initializations. ;

Step 2 (**Optimization**):

**for**  $r = 1, 2, \dots, iter$  **do**

    Update  $\mathbf{A}$  with fixed  $\mathbf{W}$  and  $\mathbf{D}$  using (5).

    Update  $\mathbf{D}$  with fixed  $\mathbf{W}$  and  $\mathbf{A}$  using (7).

    Update  $\mathbf{W}$  with fixed  $\mathbf{D}$  and  $\mathbf{A}$  using (11).

    If  $|\mathbf{W}^r - \mathbf{W}^{r-1}| < \epsilon$  and  $r > 2$ , go to Step 3.

**end**

Step 3 (**Output**):

Output the projection matrix  $\mathbf{W}$ , dictionary  $\mathbf{D}$ , and coefficient matrix  $\mathbf{A}$ .

## 4 COUPLED-SLBFLE

For heterogeneous face matching, we propose a coupled-SLBFLE (C-SLBFLE) method which extends our SLBFLE to the heterogeneous matching scenario. Specifically, we learn the hash functions for each modality by using a matrix factorization procedure to learn a common latent semantic space that implicitly reduces the modality gap between binary codes of different modalities. In the training stage, we extract coupled-PDVs for each heterogeneous pair which are from the same person and captured in different modalities. We employ C-SLBFLE to obtain the projection matrices (one for each modality) and a unified codebook for face representation, which is illustrated in Fig. 5.

Let  $\mathbf{X}_p = [\mathbf{x}_{p1}, \mathbf{x}_{p2}, \dots, \mathbf{x}_{pN}]$  and  $\mathbf{X}_q = [\mathbf{x}_{q1}, \mathbf{x}_{q2}, \dots, \mathbf{x}_{qN}]$  be the PDVs extracted from two ( $P$  and  $Q$ ) modalities of face image sets, where  $\mathbf{x}_{pn}$  and  $\mathbf{x}_{qn}$  are the  $n$ th PDV extracted from the first and the second modality at the same position, respectively, and  $1 \leq n \leq N$ . Our C-SLBFLE aims to seek  $K$  pairs of hash functions to map and quantize  $\mathbf{x}_{pn}$  and  $\mathbf{x}_{qn}$  into a common binary vector  $\mathbf{b}_n = [b_{n1}, \dots, b_{nK}] \in \{-1, 1\}^{K \times 1}$ . By learning a common binary vector, the modality gap between  $\mathbf{X}_p$  and  $\mathbf{X}_q$  is implicitly reduced. We are able to

learn this by a collective matrix factorization formulation [37] such that we learn latent variables  $\mathbf{U}_p \in \mathbb{R}^{d \times K}$  and  $\mathbf{U}_q \in \mathbb{R}^{d \times K}$  which learns a common semantic binary representation  $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_N] \in \{-1, 1\}^{K \times N}$ , described as below

$$\mathbf{x}_{pm} = \mathbf{U}_p \mathbf{b}_n \quad (12)$$

$$\mathbf{x}_{qn} = \mathbf{U}_q \mathbf{b}_n. \quad (13)$$

Different from [37] which learns a real-valued representation, our common semantic representation is performed in the Hamming space. Specifically, we learn the feature mapping matrices by minimizing the following quantization loss

$$\min_{\mathbf{W}_p, \mathbf{W}_q} \|\mathbf{B} - \mathbf{W}_p^\top \mathbf{X}_p\|_F^2 + \|\mathbf{B} - \mathbf{W}_q^\top \mathbf{X}_q\|_F^2, \quad (14)$$

where  $\mathbf{W}_p \in \mathbb{R}^{d \times K}$  and  $\mathbf{W}_q \in \mathbb{R}^{d \times K}$  are the projection matrices for the modality  $P$  and  $Q$ , respectively. In learning these projection matrices, we assume that  $\mathbf{W}_p$  and  $\mathbf{W}_q$  contain a shared information which maximizes the similar properties between the two modalities, and modality-specific properties which are only unique to each modality. Hence, we define each projection matrix as  $\mathbf{W}_p = [\mathbf{W}_c, \mathbf{W}_s^{(p)}]$  and  $\mathbf{W}_q = [\mathbf{W}_c, \mathbf{W}_s^{(q)}]$ , where  $\mathbf{W}_c \in \mathbb{R}^{d \times K_c}$  is the shared projection across the two modalities,  $\mathbf{W}_s^{(p)} \in \mathbb{R}^{d \times K_s}$  and  $\mathbf{W}_s^{(q)} \in \mathbb{R}^{d \times K_s}$  are the individual specific projections for the modality  $P$  and  $Q$ , respectively. The ratio between the dimension of each is determined by the hyperparameter  $\eta = \frac{K_s}{K_s + K_c}$ , where the total projection length is  $K = K_s + K_c$ .

To perform the binary embedding from PDV to the common semantic representation during testing phase, we simply binarize the feature projection as follows:

$$\mathbf{b}_{pm} = \text{sgn}(\mathbf{W}_p^\top \mathbf{x}_{pm}) \quad (15)$$

$$\mathbf{b}_{qn} = \text{sgn}(\mathbf{W}_q^\top \mathbf{x}_{qn}). \quad (16)$$

To learn the couple latent variables and the common binary code, we combine (14)-(16) to our SLBFLE formulation and present the following optimization objective function

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{W}_p, \mathbf{W}_q, \mathbf{U}_p, \mathbf{U}_q, \mathbf{D}, \mathbf{A}} J &= J_1 + \lambda_1 J_2 + \lambda_2 J_3 + \mu R(\cdot) \\ &= \|\mathbf{B} - \mathbf{D}\mathbf{A}\|_F^2 + \gamma \|\mathbf{A}\|_1 \\ &\quad + \lambda_1 (\|\mathbf{X}_p - \mathbf{U}_p \mathbf{B}\|_F^2 + \|\mathbf{X}_q - \mathbf{U}_q \mathbf{B}\|_F^2) \\ &\quad + \lambda_2 (\|\mathbf{B} - \mathbf{W}_p^\top \mathbf{X}_p\|_F^2 + \|\mathbf{B} - \mathbf{W}_q^\top \mathbf{X}_q\|_F^2) \\ &\quad + \mu R(\mathbf{U}_p, \mathbf{U}_q, \mathbf{W}_p, \mathbf{W}_q) \end{aligned} \quad (17)$$

subject to  $\mathbf{B}\mathbf{B}^\top = \mathbf{N}\mathbf{I}^{k \times k}$

$$\mathbf{W}_p = [\mathbf{W}_s^{(p)}, \mathbf{W}_c], \quad \mathbf{W}_q = [\mathbf{W}_s^{(q)}, \mathbf{W}_c]$$

$$\mathbf{B} = \{-1, 1\}^{n \times k},$$

The objective of  $J_1$  is to learn a dictionary  $\mathbf{D}$  over the binary codes where each binary vector can be sparsely reconstructed by  $\mathbf{A}$ . The aim of  $J_2$  is to learn the latent variables that minimizes the modality gap of the two modalities.  $J_3$  minimizes the quantization loss between the learned binary code and real-value code obtained from the shared structure feature mapping. Similar, to SLBFLE, we also have a constraint that ensures that each projection vector leads to

independent and uncorrelated binary vectors through maximizing the variance.  $R(\cdot) = \|\cdot\|_F^2$  is regularization term to avoid the over-fitting with  $\mu$  as a balancing parameter.  $\lambda_1$  and  $\lambda_2$  balances the weight of the different criterions in the overall optimization function. Our optimization function can also be solved through an iterative procedure such that we fix the other variables and solve the one variable. The detailed procedure is described as follows:

*Learning  $\mathbf{A}$  and  $\mathbf{D}$  with other fixed parameters:* Similar to SLBFLE, we solve the optimization problem in (6) and (7) by using the feature sign search and conjugate gradient descent for  $\mathbf{A}$  and  $\mathbf{D}$ , respectively.

*Learning  $\mathbf{U}_t$  where  $t = \{p, q\}$  with other fixed parameters:* By fixing other parameters, we obtain  $\mathbf{U}_t$  by solving the following regularized least squares problem

$$\min_{\mathbf{U}_t} J(\mathbf{U}_t) = \lambda_1 (\|\mathbf{X}_t - \mathbf{U}_t \mathbf{B}\|_F^2) + \mu \|\mathbf{U}_t\|_F^2, \quad (18)$$

which can be easily calculated as in a closed form solution as follows:

$$\mathbf{U}_p = \mathbf{X}_p \mathbf{B}^\top \left( \mathbf{B}\mathbf{B}^\top + \frac{\mu}{\lambda_1} \mathbf{I} \right)^{-1} \quad (19)$$

$$\mathbf{U}_q = \mathbf{X}_q \mathbf{B}^\top \left( \mathbf{B}\mathbf{B}^\top + \frac{\mu}{\lambda_1} \mathbf{I} \right)^{-1}, \quad (20)$$

where  $\mathbf{I}$  is and identity matrix.

*Learning  $\mathbf{B}$  with other fixed parameters:* By fixing other parameters, we see that (17) is non-differentiable due to the employed  $\text{sgn}$  function. Originally, SLBFLE relaxes the binary code to be the signed magnitude,  $\mathbf{B} = \mathbf{W}^\top \mathbf{X}$ . However, we cannot relax the common binary code as a signed magnitude from either modality  $\mathbf{X}_p$  or  $\mathbf{X}_q$ . In this work, we relax  $\mathbf{B}$  as its own unique real-value code and then perform binarization. The solution is obtained as follows:

$$\begin{aligned} \mathbf{B} &= \text{sgn} \left( \lambda_1 \left( \sum_{t=p,q} \mathbf{U}_t^\top \mathbf{U}_t \right) + \lambda_3 \mathbf{I} \right)^{-1} \\ &\quad \times \left( \mathbf{D}\mathbf{A} + \sum_{t=p,q} (\lambda_1 \mathbf{U}_t^\top + \lambda_2 \mathbf{W}_t^\top) \mathbf{X}_t \right), \end{aligned} \quad (21)$$

where  $\lambda_3$  is a new parameter that penalizes the independent bit constraint which is similarly relaxed by maximizing the variance instead.

*Learning  $\mathbf{W}_s^{(t)}$  where  $t = \{p, q\}$  with other fixed parameters:* To learn the modal-specific feature mapping matrices, we fix other parameters and obtain  $\mathbf{W}_s^{(t)}$  as follows:

$$\mathbf{W}_s^{(p)} = \left( \mathbf{X}_p \mathbf{X}_p^\top + \frac{\mu}{\lambda_2} \mathbf{I} \right)^{-1} \mathbf{X}_p \mathbf{B}_s^\top \quad (22)$$

$$\mathbf{W}_s^{(q)} = \left( \mathbf{X}_q \mathbf{X}_q^\top + \frac{\mu}{\lambda_2} \mathbf{I} \right)^{-1} \mathbf{X}_q \mathbf{B}_s^\top. \quad (23)$$

Differently, the full binary code is the concatenation of a set of binary codes,  $\mathbf{B} = [\mathbf{B}_s; \mathbf{B}_c]$ . The first set  $\mathbf{B}_s$  is obtained from the modal-specific feature mapping, and the other set  $\mathbf{B}_c$  is obtained from the common feature mapping. Here, only the modal-specific binary codes are used in the solution.

*Learning  $\mathbf{W}_c$  with other fixed parameters:* To learn the common feature mapping matrix, we fix other parameters and obtain  $\mathbf{W}_c$  by solving the following optimization problem

$$\min_{\mathbf{W}_c} J(\mathbf{W}_c) = \lambda_2 (\|\mathbf{B}_c - \mathbf{W}_c \mathbf{X}_p\|_F^2 + \|\mathbf{B}_c - \mathbf{W}_c \mathbf{X}_q\|_F^2) + \mu \|\mathbf{W}_c\|_F^2. \quad (24)$$

We simplify this problem by combining the two terms into a single regularized least square problem which can be solved as follows:

$$\mathbf{W}_c = \left( \mathbf{X}_c \mathbf{X}_c^\top + \frac{\mu}{\lambda_2} \mathbf{I} \right)^{-1} \mathbf{X}_c \tilde{\mathbf{B}}_c^\top. \quad (25)$$

where  $\tilde{\mathbf{B}}_c = [\mathbf{B}_p; \mathbf{B}_q]$  and  $\mathbf{X}_c = [\mathbf{X}_p; \mathbf{X}_q]$ .

We repeat the above steps until the algorithm converges. Algorithm 2 summarizes the detailed procedure of the proposed C-SLBFLE method.

---

#### Algorithm 2. C-SLBFLE

---

**Input:** Training set  $\mathbf{X}_p = [\mathbf{x}_{p1}, \mathbf{x}_{p2}, \dots, \mathbf{x}_{pN}]$ ,  $\mathbf{X}_q = [\mathbf{x}_{q1}, \mathbf{x}_{q2}, \dots, \mathbf{x}_{qN}]$ , iteration number  $iter$ , parameters  $\lambda_1, \lambda_2, \lambda_3, \mu$ , binary code length  $K$ , and shared structure ratio  $\eta$ .

**Output:** Shared Projection matrices  $\mathbf{W}_p$  and  $\mathbf{W}_q$ , dictionary  $\mathbf{D}$ .

Step 1 (**Initialization**):

1.1 Initialize  $\mathbf{W}_p$  and  $\mathbf{W}_q$  as the top  $K$  eigenvectors of obtained from CCA corresponding to the  $K$  largest eigenvalues.

1.2 Initialize  $\mathbf{U}_p, \mathbf{U}_q, \mathbf{D}$  and  $\mathbf{A}$  with arbitrary initializations.

Step 2 (**Optimization**):

**for**  $r = 1, 2, \dots, iter$  **do**

    Update  $\mathbf{A}$  using (6).

    Update  $\mathbf{D}$  using (7).

    Update  $\mathbf{U}_t$  using (19) and (20)

    Update  $\mathbf{B}$  using (21)

    Update  $\mathbf{W}_s^{(t)}$  using (22) and (23)

    Update  $\mathbf{W}_c$  using (25)

    If  $|\mathbf{W}_p^r - \mathbf{W}_p^{r-1}| < \epsilon, |\mathbf{W}_q^r - \mathbf{W}_q^{r-1}| < \epsilon$  and  $r > 2$ , go to Step 3.

**end**

Step 3 (**Output**):

Output the projection matrix  $\mathbf{W}_p = [\mathbf{W}_s^{(p)}; \mathbf{W}_c]$ ,

$\mathbf{W}_q = [\mathbf{W}_s^{(q)}; \mathbf{W}_c]$ , and the dictionary  $\mathbf{D}$ .

---

## 5 EXPERIMENTS

We conducted face recognition experiments to evaluate our SLBFLE and C-SLBFLE methods on six widely used face datasets. Specifically, we implemented our SLBFLE method on the FERET, LFW, YTF and PaSC datasets in face recognition and verification, and evaluated our C-SLBFLE method on the CASIA NIR-VIS 2.0 and Multi-PIE dataset for heterogeneous face recognition. The following describes the details of the experiments and results.

### 5.1 Results on LFW

The LFW dataset [38] contains 13,233 images from 5,749 persons. Facial images in this dataset were collected from the web, so that there are large intra-class variations in pose, illumination and expression because these images are captured in wild conditions. In our experiments, we evaluated

TABLE 1  
Mean Verification Rate (VR) (%) and Area Under ROC (AUC) Comparison with State-of-the-Art Face Descriptors on LFW with the Unsupervised Setting

Method	VR	AUC
LBP [40]	69.45	75.47
SIFT [40]	64.10	54.07
LARK [41]	72.23	78.30
POEM [42]	75.22	-
LHS [43]	73.40	81.07
MRF-MLBP [44]	80.08	89.94
PEM (LBP) [45]	81.10	-
PEM (SIFT) [45]	81.38	-
DFD [15]	84.02	-
CBFD (combine) [31]	-	90.91
High-dim LBP [46]	84.08	-
PAF [47]	<b>87.77</b>	94.05
MRF-Fusion-CSKDA [44]	-	<b>98.94</b>
Spartans [48]	-	94.24
LBPNet [49]	-	94.04
SLBFLE (R=2)	82.02	88.95
SLBFLE (R=3)	84.08	90.46
SLBFLE (R=4)	84.18	90.53
SLBFLE (R=2+3+4)	<b>85.62</b>	<b>92.00</b>

our proposed method with the unsupervised setting and the image-restricted with label-free outside data setting. We followed the standard evaluation protocol on the ‘‘View 2’’ dataset which includes 3,000 matched pairs and 3,000 mismatched pairs and is divided into 10 folds, where each fold consists of 300 matched (positive) pairs and 300 mismatched (negative) pairs. We used the aligned LFW-a dataset<sup>2</sup> for our evaluation, where each face image in LFW was aligned and cropped into  $128 \times 128$  to remove the background information. We learned feature representation with our proposed SLBFLE. Specifically, each PDV was first projected into a  $K$ -bit binary codes with the learned projection  $\mathbf{W}$  and then encoded as a feature with the learned dictionary  $\mathbf{D}$ . The parameters  $\lambda_1, \lambda_2$  were empirically tuned as 0.001 and 0.01, respectively, by using a cross-validation strategy on the ‘‘View 1’’ subset of the LFW dataset. We tested our method with different neighborhood radius sizes ( $R$  is set as 2, 3 and 4), which yields a 24-, 48-, and 80-dimensional PDV, respectively. We further applied the whitened PCA (WPCA) method to project each sample into a 500-dimensional feature vector to reduce the redundancy. For the unsupervised setting, the nearest neighbor classifier with the cosine similarity was used for face verification. For the image-restricted with label-free outside data setting, we used the discriminative deep metric learning (DDML) [39] method to learn discriminative similarity measure function for face verification.

#### 5.1.1 Comparison with the State-of-the-Art Methods

Table 1 tabulates the average verification rate and Fig. 6 shows the ROC curve of our SLBFLE on LFW with the unsupervised setting, as well as those of the state-of-the-art face feature descriptors. We see that SLBFLE achieves better performance than existing hand-crafted feature descriptors such as LARK and PEM, and obtains very competitive

2. Available: <http://www.openai.com/home/hassner/data/lfwa/>.

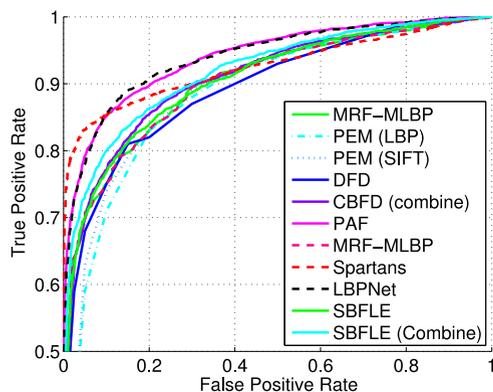


Fig. 6. ROC curve of different face descriptors on LFW with the unsupervised setting.

performance with the existing learning-based feature descriptors such as DFD.<sup>3</sup> Moreover, the performance of SLBFLE can be further improved when multiple PDVs with different neighboring sizes are combined. Specifically, we combined the SLBFLE features extracted from varying neighboring radius at the decision level. For the unsupervised setting, we took the average of the similarity scores of the face image pairs obtained from the SLBFLE features of varying neighboring radius. For the restricted setting, we first employed DDML [39] for each SLBFLE feature type to obtain the similarity score for each face image pair. Then, these similarity scores from different features were concatenated into a single score vector which was trained using linear SVM<sup>4</sup> in the training set. Finally, we used the SVM model to learn the final verification score and applied it for testing samples.

Table 2 tabulates the average verification rate and Fig. 7 shows the ROC curve of our SLBFLE on LFW with the image-restricted with label-free outside data setting, as well as those of the state-of-the-art face verification methods. We see that our SLBFLE method with multiple PDVs extracted from different neighboring sizes outperforms most of the current state-of-the-art methods. Moreover, the performance of SLBFLE can be further boosted when it is combined with several other existing hand-crafted feature descriptors. Similarly, we combined our SLBFLE with 5 other existing feature descriptors including the Sparse SIFT [39], Dense SIFT [39], low-dimensional LBP [39], HOG [39], and high-dimensional LBP [46]. Then, these similarity scores from different features are concatenated into a single score vector which is trained using Linear SVM in the training set. Finally, we used the SVM model to learn the final verification score and applied it for testing samples. By doing so, the mean verification rate can be further improved by 2.79 percent.

While the methods in [44] and [50] show better performance than our method, it is important to note that the method in [44] combined three hand-crafted features (MLBP, MLPQ and MBSIF), which were extracted from multiple scales with multiple filters. The method in [50] also used multiple filters to extract hand-crafted features, which

3. Compared with DFD which is a supervised feature learning approach, our SLBFLE is unsupervised so that it is more convenient for practical applications.

4. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

TABLE 2  
Mean Verification Rate and the Standard Error (%) Comparison with State-of-the-Art Face Verification Methods on LFW with the Image-Restricted with Label-Free Outside Data Setting

Method	Accuracy
CSML+SVM [52]	88.00 ± 0.37
High-Throughput BIF [53]	88.13 ± 0.58
LARK supervised [41]	85.10 ± 0.59
DML-eig combined [54]	85.65 ± 0.56
Covolutional DBN [19]	87.77 ± 0.62
STFRD+PMML [55]	89.35 ± 0.50
PAF [47]	87.77 ± 0.51
Sub-SML [56]	89.90 ± 0.38
VMRS [57]	91.10 ± 0.59
DDML [39]	90.68 ± 1.41
LM3L [58]	89.57 ± 1.53
Hybrid on LFW3D [59]	85.63 ± 0.53
Sub-SML + Hybrid on LFW3D [59]	91.65 ± 1.04
HPEN + HD-LBP + DDML [60]	92.57 ± 0.36
HPEN + HD-Gabor + DDML [60]	92.80 ± 0.47
CBFD (combine)	92.62 ± 1.08
Spartans [51]	89.69 ± 0.36
MSBSIF-SIEDA [50]	94.63 ± 0.95
TSML with OCLBP [61]	87.10 ± 0.43
TSML with feature fusion [61]	89.80 ± 0.47
SLBFLE (R=2)	85.62 ± 1.41
SLBFLE (R=3)	86.57 ± 1.65
SLBFLE (R=4)	87.45 ± 1.28
SLBFLE (R=2+3+4)	90.18 ± 1.89
SLBFLE (All combined)	<b>92.97 ± 1.20</b>

is computationally expensive. Similarly, LBP-Net [49] also used multiple types of LBP operators which were obtained by multiple filters. While the method in [51] achieved promising recognition performance, it requires high-dimensional weighted LBP features extracted from multiple scales. Differently, our SLBFLE can achieve reasonably good performance by using only three scales ( $R = 2, 3, 4$ ).

## 5.2 Results on YTF

The YTF dataset [62] contains 3,425 videos of 1,595 different persons collected from the YouTube website. There are large variations in pose, illumination, and expression in each video, and the average length of each video clip is 181.3 frames. In our experiments, we followed the standard evaluation protocol [62] and tested our method for unconstrained face

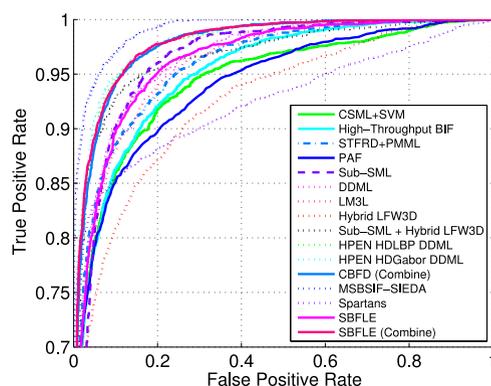


Fig. 7. ROC curve of different face verification methods on LFW with the image-restricted with label-free outside data setting.

TABLE 3  
Comparisons of the Mean Verification Rate and Standard Error (%) with State-of-the-Art Learning-Based Face Descriptors on YTF Under the Image-Restricted Setting

Method	Accuracy
LBP [5]	75.86 ± 1.42
FPLBP [64]	73.58 ± 1.62
CSLBP [65]	73.70 ± 1.63
LE [13]	69.72 ± 2.06
DFD [15]	78.10 ± 0.94
SLBFLE (R=2)	80.35 ± 0.84
SLBFLE (R=3)	81.24 ± 1.32
SLBFLE (R=4)	82.36 ± 1.01
SLBFLE (R=2+3+4)	<b>82.88 ± 1.01</b>

verification with 5,000 video pairs. These pairs are equally divided into 10 folds, and each fold has 250 intra-personal pairs and 250 inter-personal pairs. For each video clip, we first learned feature representation for each frame by using our SLBFLE method, and then averaged all the feature vectors within one video clip to form a mean feature vector in our experiments because all face images have been aligned by the detected facial landmarks. Lastly, we used WPCA to project each mean vector into a 500-dimensional feature vector. Similarly, we also used the DDML method for face verification with the image-restricted setting.

Table 3 tabulates the average verification rate of our method and three state-of-the-art learning-based face descriptors on YTF. We see that our method outperforms these state-of-the-art methods with the smallest gain of 2.25 percent in terms of the mean verification rate. Moreover, the performance of SLBFLE can be further improved when multiple PDVs with different neighboring sizes are combined.

Table 4 tabulates the average verification rates and Fig. 8 shows the ROC curves of our method and state-of-the-art face verification methods on YTF. We see that our method achieves very competitive performance with state-of-the-art methods. By combining our method with different values of  $R$  and binary pattern descriptors (LBP, CSLBP and FPLBP), we achieve very competitive results with the other compared methods such as Eigen-Pep [63] and DDML [39]. Moreover, we obtain better performance than the DeepFace method [22] with a gain of 2 percent in accuracy by combining our method at different values of  $R$  and the convolutional neural network (CNN) features which were trained with fewer labeled samples. For each face image, we extracted the CNN feature by using the pre-trained VGG-face network in [23]<sup>5</sup> and combined it with our SLBFLE feature at the decision level. Specifically, for the unsupervised setting, we took the average of the similarity scores of the face image pairs obtained from the CNN features and SLBFLE features before using them for computing the ROC curve. For the restricted setting, we first employed DDML [39] for the CNN and SLBFLE features individually to obtain the similarity scores for each face image pair, respectively. The similarity scores for different features were then concatenated to a single score vector which was

TABLE 4  
Comparisons of the Mean Verification Rate and Standard Error (%) with the State-of-the-Art Face Verification Methods on YTF Under the Image-Restricted Setting

Method	Accuracy
MBGS(LBP) [62]	76.4 ± 1.8
MBGS+SVM (LBP) [66]	78.9 ± 1.9
APEM(fusion) [45]	79.1 ± 1.5
STFRD+PMML [55]	79.5 ± 2.5
VSOFF+OSS [67]	79.7 ± 1.8
DDML (LBP) [39]	81.3 ± 1.6
DDML (combined) [39]	82.3 ± 1.5
EigenPEP [63]	84.8 ± 1.4
LM3L [39]	81.3 ± 1.2
DeepFace [22]	<b>91.4 ± 1.1</b>
SLBFLE (R=2+3+4)	82.9 ± 1.0
SLBFLE (R=2+3+4 + LBP + CSLBP+ FPLBP)	83.4 ± 1.0
SLBFLE (R=2+3+4 + CNN)	<b>93.4 ± 1.0</b>

trained by using linear SVM in the training set. Finally, we used the SVM model to learn the final verification score and applied it for testing samples.

### 5.3 Results on FERET

The FERET dataset consists of 13,539 face images of 1,565 subjects who are diverse across age, gender, and ethnicity. We followed the standard FERET evaluation protocol [68], where six sets including the *training*, *fa*, *fb*, *fc*, *dup1*, and *dup2* were constructed for experiment, respectively. All face images were scaled and cropped into  $128 \times 128$  pixels according to the provided eye coordinates. We performed feature learning on the generic *training* set, and applied the learned projection and dictionary matrix on the other five sets for feature extraction. Finally, we take *fa* as the gallery set and the other four sets as the probe sets. We followed the same parameter settings which were tuned on LFW. We applied WPCA to project each sample into a 1196-dimensional feature vector and applied the nearest neighbor classifier with the cosine similarity for face identification.

Table 5 tabulates the rank-one identification rate of our method, as well as the state-of-the-art feature descriptors on the FERET dataset. We see that our SLBFLE achieves the competitive recognition rate on all four subsets. Specifically, SLBFLE achieves much better performance than hand-crafted

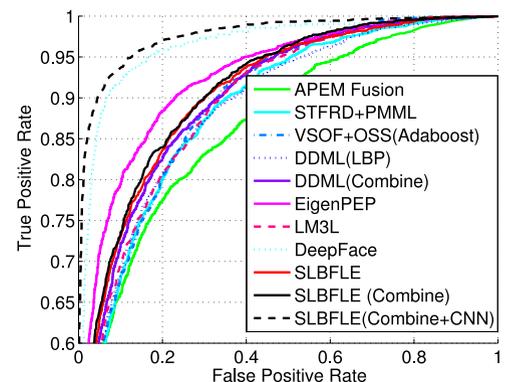


Fig. 8. ROC curve of different face verification methods on YTF under the image-restricted setting.

5. [http://www.robots.ox.ac.uk/~vgg/software/vgg\\_face/](http://www.robots.ox.ac.uk/~vgg/software/vgg_face/).

TABLE 5  
Rank-One Recognition Rates (%) Comparison with  
State-of-the-Art Feature Descriptors with the  
Standard FERET Evaluation Protocol

Method	fb	fc	dup1	dup2
LBP [5]	93.0	51.0	61.0	50.0
LGBP [6]	94.0	97.0	68.0	53.0
HGGP [8]	97.6	98.9	77.7	76.1
LDP [9]	94.0	83.0	62.0	53.0
GV-LBP-TOP [10]	98.4	99.0	82.0	81.6
GV-LBP [10]	98.1	98.5	80.9	81.2
LQP [14]	99.8	94.3	85.5	78.6
POEM [11]	97.0	95.0	77.6	76.2
s-POEM [69]	99.4	<b>100.0</b>	91.7	90.2
DFD [15]	99.4	<b>100.0</b>	91.8	92.3
CBFD [31]	99.8	<b>100.0</b>	93.5	<b>93.2</b>
SLBFLE (R=2)	99.7	99.7	89.9	80.0
SLBFLE (R=3)	<b>99.9</b>	<b>100.0</b>	94.5	90.9
SLBFLE (R=4)	<b>99.9</b>	<b>100.0</b>	<b>95.2</b>	<b>92.7</b>

feature descriptors such as HGGP, GV-LBP-TOP and GV-LBP. This is because our SLBFLE is a data-adaptive feature representation method. Compared with the recently proposed learning-based feature representation methods such as DFD and CBFD, our SLBFLE is a binary code based feature descriptor which can demonstrate stronger robustness to local variations. Hence, higher recognition rates can be obtained.

#### 5.4 Results on PaSC

The PaSC dataset [70], [71] contains 9,376 images and 2,802 videos from 293 people. These images and videos were captured in different viewpoints, poses and distances from the camera, which make it challenging for face recognition. We performed still-video face matching experiments on the PaSC dataset where we were provided with 1,401 handheld videos as the query set, and 4,688 still images as the target set. Each face image from both sets was aligned with the provided eye coordinates and cropped into a face region with the size of  $128 \times 128$ . We extracted the feature for each image from the query and target sets and projected it into a 100-dimensional feature space with WPCA and LDA. Following the standard evaluation protocol in [71], we compared face images in the query set to the target set and obtained a  $1,401 \times 4,688$  similarity matrix by computing the cosine distances between samples. For the handheld video set, we only randomly selected 20 face representations for each video and selected the max similarity score. We compared our method with the LRPCA baseline and other popular face matching methods [71], [72]. Table 6 shows the verification rates and Fig. 9 shows the ROC curves of different methods. As can be seen, our proposed SLBFLE outperforms the compared methods, where the minimal improvement of the verification rate is 10 percent at 0.01 FAR.

#### 5.5 Results on CASIA NIR-VIS 2.0

The CASIA NIR-VIS 2.0 dataset [76] was used for the heterogenous face matching to evaluate our C-SLBFLE method. This dataset consists of 725 identities, where each identity consists of VIS (Visual) and NIR (Near Infrared) images having a range of 1 to 22 and 5 to 50 face images,

TABLE 6  
The Verification Rate at the 1.0 Percent FAR of Different  
Methods on the PaSC Dataset Still-Video Face  
Recognition Experiments

Method	Verification rate
LBP-SIFT-WPCA-SILD [73]	0.23
ISV-GMM [74]	0.11
PLDA-WPCA-LLR [75]	0.26
LRPCA Baseline [70]	0.10
Eigen-PEP [63]	0.24
Hierarchical-PEP [72]	0.32
SLBFLE (R=2)	0.25
SLBFLE (R=3)	0.34
SLBFLE (R=4)	<b>0.42</b>

respectively. This dataset is the largest one for the VIS-NIR heterogenous face matching problem. We used the pre-defined 10-fold split and evaluation protocol from the dataset providers using the VIS and NIR images as gallery and probe set, respectively.

For the pre-processing, we first cropped and aligned each face image into a  $128 \times 128$  image size based on the obtained eye coordinates. Then, we performed filtering through the Difference of Gaussians (DoG) filter to reduce the variations from each modality. Because each identity does not contain similar number of VIS and NIR images, we randomly selected VIS-NIR pairs having the same identity to be passed through the C-SLBFLE during learning. In our implementation, the PDVs were extracted using  $K = 20$  and an  $\eta = 0.25$ . We set  $\lambda_1 = 0.02$ ,  $\lambda_2 = 0.005$  and  $\lambda_3 = 0.001$  based on cross-validation experiment. Having obtained the C-SLBFLE face representation from the learned projection matrices and codebook, each representation was then projected into a 1000-dimensional feature vector through WPCA. The nearest neighbor classifier was then used to retrieve the final match where we used the cosine distance as the similarity measure between two representations. Moreover, the LDA method was further employed to extract more discriminative information to improve the recognition performance.

We evaluated the heterogenous face matching procedure based on the rank-one recognition rates and verification rate at a false acceptance rate (FAR) of 0.1 and 0.01 percent. We compared our method with several subspace methods (CCA, PLS, CDFE, MvDA), hand-crafted feature descriptors

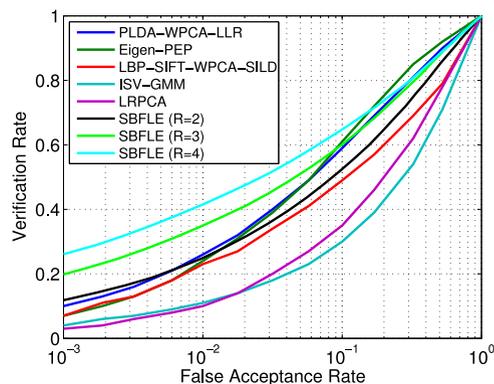


Fig. 9. ROC curves of different feature descriptors on the PaSC dataset in the unsupervised setting.

TABLE 7  
Performance Comparison on the CASIA NIR-VIS 2.0 Dataset, Where VR Is (twice) the Mean Verification Rate When the FAR Is Set to 0.1 and 1.0 Percent

Method	Rank-one Acc.± std%	VR@ 0.1(%)	VR@ 1(%)
CCA [77]	28.5 ± 3.4	10.8	30.7
PLS [78]	17.7 ± 1.9	2.3	9.5
CDFE [79]	27.9 ± 2.9	6.9	23.3
MvDA [80]	41.6 ± 4.1	19.2	42.8
LCFS [81]	35.4 ± 2.8	16.7	35.7
GMLDA [82]	23.7 ± 1.4	5.1	16.6
GMMFA [82]	24.8 ± 1.1	7.6	19.5
LBP [5]	35.4 ± 2.7	4.2	31.8
LTP [83]	35.1 ± 2.2	8.2	34.7
TP-LBP [64]	36.2 ± 1.6	3.7	12.9
FP-LBP [64]	23.2 ± 1.0	1.7	9.0
LPQ [84]	47.5 ± 0.9	4.0	17.2
SIFT [85]	49.1 ± 2.3	14.3	40.8
PCA+Sym+HCA [45]	23.7 ± 1.9	19.3	-
DSIFT+PCA+LDA [86]	73.3 ± 1.1	-	-
C-DFD [15]	65.8 ± 1.6	46.2	61.9
CDFL [26]	71.5 ± 1.4	55.1	67.7
Gabor+RBM remove 11PCs [25]	<b>86.2 ± 1.0</b>	81.3	-
C-CBFD [31]	81.8 ± 2.3	47.3	75.3
CNN + LDML [24]	85.9 ± 1.0	78.0	-
Recon. + UDP(DLBP) [48]	78.5 ± 1.8	<b>85.8</b>	-
SLBFLE (R=2)	75.8 ± 2.6	38.5	69.8
SLBFLE (R=3)	79.0 ± 2.9	43.1	72.5
SLBFLE (R=4)	80.6 ± 2.4	45.5	74.5
C-SLBFLE (R=2)	79.4 ± 3.2	41.9	72.1
C-SLBFLE (R=3)	85.9 ± 2.1	50.5	78.5
C-SLBFLE (R=4)	<b>86.9 ± 2.2</b>	<b>53.0</b>	<b>80.3</b>

(LBP, LTP, SIFT), and feature learning methods (C-DFD, CDFL, C-CBFD). Table 7 shows the results of our method in comparison with most state-of-the-art heterogeneous face recognition methods. We see that our C-SLBFLE achieves better performance than the hand-crafted feature descriptors because our method performs coupled feature learning from a training set such that the coupled-PDVs can be projected as similar as possible despite the modality differences. Our C-SLBFLE also beats the subspace methods because our method reduces the modality gap at the feature level. Fig. 10 shows the ROC curve for our model compared to other feature descriptors. Moreover, our C-SLBFLE method outperforms other feature learning methods.

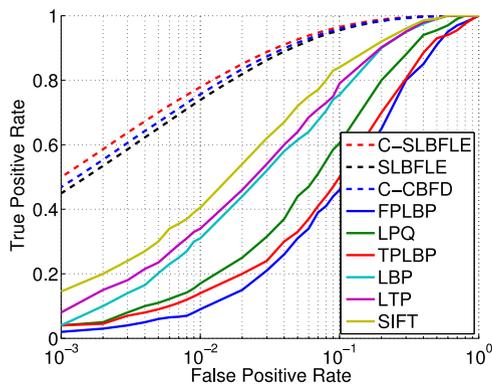


Fig. 10. ROC curves of different methods on the CASIA NIR-VIS 2.0 dataset.

TABLE 8  
Performance Comparison of C-SLBFLE ( $R = 4$ ) the CASIA NIR-VIS 2.0 Dataset at Varying  $\eta$ , Where VR Is (twice) the Mean Verification Rate When the FAR Is Set to 0.1 and 1.0 Percent

Method	Rank-one Acc ± std (%)	VR @FAR=0.1(%)	VR @FAR=1(%)
$\eta = 0$	85.0 ± 2.6	50.2	78.5
$\eta = 0.25$	<b>86.9 ± 2.2</b>	<b>53.0</b>	<b>80.3</b>
$\eta = 0.5$	83.5 ± 3.0	47.8	76.0
$\eta = 0.8$	82.1 ± 2.6	45.8	74.6
$\eta = 1$	80.7 ± 2.5	42.4	73.4

Specifically, our method performed better than our previous C-CBFD [31] which also employs coupled feature learning by maximizing the correlation of matching binary vectors. This is because our C-SLBFLE method performs joint feature learning and codebook encoding which enhances the robustness of each part. We also see that adding the collective matrix factorization term to learn latent variables that implicitly reduces the modality gap boosted the performance by 5 percent in the rank-one accuracy. Our method is also very competitive to the best results in [25]. It is important to note that [25] performed an additional landmark detection step to obtain representative Gabor features for learning, while our C-SLBFLE method performed dense feature extraction. While [48] also achieved promising performance, their method performed homogeneous data reconstruction by using cross-spectral joint dictionary learning before extracting hand-crafted descriptors, which increases the computational cost of the recognition procedure.

To show the effectiveness of the shared structure learning, we also conducted experiments by only learning specific projections for each modality at varying values of  $\eta$ . Here,  $\eta = 1$  leads to the case of learning only modality specific projection matrices:  $\mathbf{W}_p = \mathbf{W}_s^{(p)}$  and  $\mathbf{W}_q = \mathbf{W}_s^{(q)}$ , while  $\eta = 0$  leads to the case of learning only a common projection matrix for both modality:  $\mathbf{W}_p = \mathbf{W}_q = \mathbf{W}_c$ . Table 8 shows the performance of our C-SLBFLE at varying  $\eta$  at  $R = 4$ . It can be seen that at  $\eta = 0.25$  shows the best result across all values of  $\eta$ , which indicates that both the common and specific properties provide useful information in order to learn the unified binary code. Most particularly, there is a performance drop at  $\eta = 1$  which indicates that learning modality-specific information contributes less to the overall performance of our method.

## 5.6 Results on Multi-PIE

The Multi-PIE dataset [87] was used for face matching on varying poses to evaluate our C-SLBFLE method to show its effectiveness of another heterogeneous face recognition applications. This dataset consists of 337 identities at 7 poses ( $-45^\circ, -30^\circ, -15^\circ, 0^\circ, +15^\circ, +30^\circ, +45^\circ$ ) captured in four sessions with varying illumination and expression. Similar to [88], we used the neutral expression with frontal illumination for the evaluations where the first 200 identities were used as training, and the remaining 137 identities for testing. Specifically, the remaining 137 captured in the first session of frontal pose were used as the gallery set and other face images of varying pose were used as the probe set. We

TABLE 9  
The Recognition Rate (%) of Different Heterogeneous Face Recognition Methods on the MultiPIE Dataset at Varying Poses with Neutral Expression and Frontal Illumination

Method	-45°	-30°	-15°	+15°	+30°	+45°	Avg
SPAЕ [73]	84.9	92.6	96.3	95.7	94.3	84.4	91.4
MvDA [80]	80.1	94.5	<b>100.0</b>	<b>100.0</b>	<b>96.6</b>	87.8	93.2
DFD [15]	54.7	94.9	<b>100.0</b>	<b>100.0</b>	92.7	<b>91.2</b>	88.9
SLBFLE	84.7	96.6	<b>100.0</b>	99.3	95.6	86.1	93.7
C-SLBFLE	<b>85.4</b>	<b>98.5</b>	<b>100.0</b>	<b>100.0</b>	96.4	87.6	<b>94.7</b>

cropped each image using the MTCNN [89] and resized it into  $128 \times 128$  for feature extraction and matching.

Table 9 shows the recognition performance of different methods. We used a similar set-up as the CASIA VIS-NIR set-up with  $R = 4$  and  $K = 1,000$ . Overall, our C-SLBFLE method is very competitive to other compared methods.

## 5.7 Analysis

### 5.7.1 Performance Analysis of Different Components in SLBFLE

We investigated three other baselines (SLBFLE1, SLBFLE2 and SLBFLE3) of our SLBFLE to show the contribution of each term of the objective function. SLBFLE1 is a variation of our SLBFLE method with  $\lambda_1$  and  $\lambda_2$  equal to 0, SLBFLE2 is another variation of our SLBFLE method with  $\lambda_1 = 0$ , while SLBFLE3 is a variation with  $\lambda_2 = 0$ . In this experiment, we use the SLBFLE features extracted at  $R = 4$ . Table 10 shows the recognition performance of the different variations of our method on different datasets. We see that minimizing the quantization loss and applying the binary constraints during feature learning contributes to the overall performance of our method.

### 5.7.2 Global SLBFLE versus Local SLBFLE

We implemented a local learning method to further improve the SLBFLE method. Motivated by the fact that different face regions usually play different roles in face representation and the global feature learning method ignores the position information of different face regions, we implemented a local SLBFLE method where binary codes and dictionaries for different face regions individually, so that more local facial structure information for feature learning can be exploited. Specifically, we first divided each face image into  $8 \times 8$  non-overlapped regions and learned 64 projections and dictionary matrices used the SLBFLE method separately, where  $R$  was set to 4 in our experiments. Table 11 shows the recognition rates of the local and global learning strategies of the SLBFLE method on the LFW dataset with

TABLE 10  
Comparisons of the Recognition Performance of Different Variations of Our Methods, Where the Mean AUC Is Used for LFW (Unsupervised) and for PaSC, and the Mean Rank-1 Identification of All Four Subjects for FERET, Respectively

Dataset	SLBFLE1	SLBFLE2	SLBFLE3	SLBFLE
LFW	89.8	89.9	89.8	<b>90.9</b>
PaSC	36.4	37.5	37.4	<b>41.5</b>
FERET	97.5	97.6	97.6	<b>97.7</b>

TABLE 11  
Comparisons of Local and Global Learning of the SLBFLE Method in the LFW Dataset with the Unsupervised Setting

Method	Global SLBFLE	Local SLBFLE
Accuracy	84.18	<b>84.75</b>

the unsupervised setting. We see that local SLBFLE can further improve the verification rate by 0.57 percent.

### 5.7.3 Computational Time

Finally, we investigated the computational time of our method to extract the feature descriptor for one single face image. As shown in Table 12, our method is slower than the handcrafted LBP descriptor. However, it is a good trade-off when both the performance and speed are concerned. Nevertheless, our method is much faster than other feature learning method such as DFD.

## 5.8 Discussion

The above experimental results suggest the following five observations:

- 1) Our SLBFLE and C-SLBFLE are competitive feature learning methods to learn face representations for both homogeneous and heterogeneous face recognition, respectively. Our methods achieved competitive performance compared to other state-of-the-art face feature descriptors. There are two reasons: 1) one approach is a data-driven so that it is more adaptive in face representation, and 2) our approach is a one-stage learning procedure, so that some useful information for dictionary learning will not be removed in the feature learning stage.
- 2) Even if our feature learning and encoding method is unsupervised, it still has strong discriminative power because raw pixels are extracted from face images of different identities which contribute to learning a discriminative feature mapping. Moreover, the learned binary codes can well describe how pixel values change over local patches and implicitly encode important visual patterns such as edges and lines in face images. Moreover, the learned dictionary can well encode the learned binary codes so that some noisy information can be well alleviated. However, our work can be extended into a supervised version when the label information of face patches are exploited in the feature learning stage so that more discriminative power of the learned features can be obtained.

TABLE 12  
Computational Complexity for Descriptor Extraction for One Face Image

Method	Feature Dimension	Time(ms)
LBP	3,776	31.1
DFD	50,176	1452.5
SLBFLE	32,000	216.4

- 3) Our C-SLBFLE is obtained by collective matrix factorization and shared structured learning, which shows that exploiting both the common and modal-specific properties in feature learning provides more information in the overall recognition performance than learning only a single common or specific projection.
- 4) Both the quantization loss and binary constraints contributes to the final recognition performance as shown in experiments on some of the datasets.
- 5) Performing local feature learning can improve the performance of our SLBFLE method. This is because by performing SLBFLE locally, a specific codebook and projection matrices are obtained which focuses on the properties of each local region. Hence, more local information can be exploited in the learned feature descriptor and the performance is further improved.

## 6 CONCLUSION

In this paper, we have proposed a simultaneous local binary feature learning and encoding method for face recognition. To make it suitable for heterogenous face matching, we have also proposed a coupled-SLBFLE (C-SLBFLE) which performs shared structured and latent feature learning to reduce the heterogenous gap between face images of different modalities for heterogeneous face matching. Experimental results on six benchmark face databases clearly demonstrate that our method achieved better or very competitive recognition performance with the state-of-the-art face feature descriptors. How to apply our proposed methods to other computer vision applications such as object recognition and visual tracking seems to be an interesting future work.

## ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China under Grant 2016YFB1001001, the National Natural Science Foundation of China under Grants 61672306, 61572271, 61527808, 61373074 and 61373090, the National 1000 Young Talents Plan Program, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program. Part of this work was presented in [1].

## REFERENCES

- [1] J. Lu, V.E. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3721–3729.
- [2] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [3] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
- [4] W.-Y. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surveys*, vol. 35, no. 4, pp. 399–458, 2003.
- [5] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
- [6] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 786–791.
- [7] Z. Lei, S. Z. Li, R. Chu, and X. Zhu, "Face recognition with local Gabor textons," in *Proc. Int. Conf. Advances Biometrics*, 2007, pp. 49–57.
- [8] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 57–68, Jan. 2007.
- [9] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 533–544, Feb. 2010.
- [10] Z. Lei, S. Liao, M. Pietikainen, and S. Z. Li, "Face recognition by exploring information jointly in space, scale and orientation," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 247–256, Jan. 2011.
- [11] N.-S. Vu and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image matching," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1352–1365, Mar. 2012.
- [12] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
- [13] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2707–2714.
- [14] S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *Proc. British Mach. Vis. Conf.*, 2012, pp. 1–12.
- [15] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 289–302, Feb. 2014.
- [16] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1891–1898.
- [17] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.
- [18] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Conf. Advances Neural Inf. Process. Syst.*, 2007, pp. 153–160.
- [19] G. B. Huang, H. Lee, and E. G. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2518–2525.
- [20] A. Hyvärinen, J. Hurri, and P. O. Hoyer, "Independent component analysis," *Natural Image Statist.*, vol. 39, pp. 151–175, 2009.
- [21] Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng, "ICA with reconstruction cost for efficient overcomplete feature learning," in *Proc. Conf. Advances Neural Inf. Process. Syst.*, 2011, pp. 1017–1025.
- [22] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1–8.
- [23] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. British Mach. Vis. Conf.*, 2015, pp. 1–12.
- [24] S. Saxena and J. Verbeek, "Heterogeneous face recognition with CNNs," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 1–8.
- [25] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, 2015, vol. 1, pp. 1–7.
- [26] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Sec.*, vol. 10, no. 3, pp. 640–652, Mar. 2015.
- [27] J. Wang, S. Kumar, and S.-F. Chang, "Semi-supervised hashing for scalable image retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3424–3431.
- [28] Y. Gong and S. Lazebnik, "Iterative quantization: A procrustean approach to learning binary codes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 817–824.
- [29] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Conf. Advances Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.
- [30] M. Norouzi, D. Fleet, and R. Salakhutdinov, "Hamming distance metric learning," in *Proc. Conf. Advances Neural Inf. Process. Syst.*, 2012, pp. 1070–1078.
- [31] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2041–2056, Oct. 2015.

- [32] H. Zhang, J. R. Beveridge, Q. Mo, B. A. Draper, and P. J. Phillips, "Randomized intraclass-distance minimizing binary codes for face recognition," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2014, pp. 1–8.
- [33] M. Rastegari, A. Farhadi, and D. Forsyth, "Attribute discovery via predictable discriminative binary codes," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2012, pp. 876–889.
- [34] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Proc. Conf. Advances Neural Inf. Process. Syst.*, 2006, pp. 801–808.
- [35] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma, "Fast l1-minimization algorithms and an application in robust face recognition: A review," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 1849–1852.
- [36] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Programm.*, vol. 142, no. 1–2, pp. 1–38, 2013.
- [37] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2075–2082.
- [38] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, MA, USA: Tech. Rep. 07–49, 2007.
- [39] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1875–1882.
- [40] R. Verschae, J. Ruiz-del Solar, and M. Correa, "Face recognition in unconstrained environments: A comparative study," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2008, pp. 1–12.
- [41] H. J. Seo and P. Milanfar, "Face verification using the lark representation," *IEEE Trans. Inf. Forensics Sec.*, vol. 6, no. 4, pp. 1275–1286, Dec. 2011.
- [42] N.-S. Vu and A. Caplier, "Face recognition with patterns of oriented edge magnitudes," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 313–326.
- [43] G. Sharma, S. ul Hussain, and F. Jurie, "Local higher-order statistics (LHS) for texture categorization and facial analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 1–12.
- [44] S. R. Arashloo and J. Kittler, "Efficient processing of MRFS for unconstrained-pose face recognition," in *Proc. IEEE 8th Int. Conf. Biometrics: Theory Appl. Syst.*, 2013, pp. 1–8.
- [45] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3499–3506.
- [46] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3025–3032.
- [47] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3539–3545.
- [48] F. Juefei-Xu, D. K. Pal, and M. Savvides, "Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog. Workshop*, 2015, pp. 141–150.
- [49] M. Xi, L. Chen, D. Polajnar, and W. Tong, "Local binary pattern network: A deep learning approach for face recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3224–3228.
- [50] A. Ouamane, M. Bengherabi, A. Hadid, and M. Cheriet, "Side-information based exponential discriminant analysis for face verification in the wild," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, 2015, pp. 1–6.
- [51] F. Juefei-Xu, K. Luu, and M. Savvides, "Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4780–4795, Dec. 2015.
- [52] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. 13th Asian Conf. Comput. Vis.*, 2010, pp. 709–720.
- [53] D. Cox and N. Pinto, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, 2011, pp. 8–15.
- [54] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1–26, 2012.
- [55] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen, "Fusing robust face region descriptors via multiple metric learning for face recognition in the wild," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3554–3561.
- [56] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3208–3215.
- [57] O. Barkan, J. Weill, L. Wolf, and H. Aronowitz, "Fast high dimensional vector multiplication face recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1960–1967.
- [58] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan, "Large margin multi-metric learning for face and kinship verification in the wild," in *Proc. 13th Asian Conf. Comput. Vis.*, 2015, pp. 252–267.
- [59] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4295–4304.
- [60] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 787–796.
- [61] L. Zheng, K. Idrissi, C. Garcia, S. Duffner, and A. Baskurt, "Triangular similarity metric learning for face verification," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, 2015, pp. 1–7.
- [62] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 529–534.
- [63] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt, "Eigen-PEP for video face recognition," in *Proc. 13th Asian Conf. Comput. Vis.*, 2010, pp. 17–33.
- [64] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in *Proc. Eur. Conf. Comput. Vis. Workshops*, pp. 1–14, 2008.
- [65] L. Wolf, T. Hassner, and Y. Taigman, "Similarity scores based on background samples," in *Proc. 13th Asian Conf. Comput. Vis.*, 2010, pp. 88–97.
- [66] L. Wolf and N. Levy, "The SVM-minus similarity score for video face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3523–3530.
- [67] M.-V. Heydi, M.-D. Yoanna, and Z. Chai, "Volume structured ordinal features with background similarity measure for video face recognition," in *Proc. Int Conf. Biometrics*, 2013, pp. 1–6.
- [68] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
- [69] N.-S. Vu, "Exploring patterns of gradient orientations and magnitudes for face recognition," *IEEE Trans. Inf. Forensics Sec.*, vol. 8, no. 2, pp. 295–304, Feb. 2013.
- [70] J. R. Beveridge, et al., "The challenge of face recognition from digital point-and-shoot cameras," in *Proc. IEEE 8th Int. Conf. Biometrics: Theory Appl. Syst.*, 2013, pp. 1–8.
- [71] J. R. Beveridge, et al., "The IJCB 2014 PASC video face and person recognition competition," in *Proc. IEEE Int. Joint Conf. Biometrics*, 2014, pp. 1–8.
- [72] H. Li and G. Hua, "Hierarchical-pep model for real-world face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4055–4064.
- [73] M. Kan, S. Shan, D. Xu, and X. Chen, "Side-information based linear discriminant analysis for face recognition," in *Proc. British Mach. Vis. Conf.*, 2011, vol. 11, pp. 125–1.
- [74] C. McCool, R. Wallace, M. McLaren, L. El Shafey, and S. Marcel, "Session variability modelling for face authentication," *IET Biometrics*, vol. 2, no. 3, pp. 117–129, 2013.
- [75] V. Struc, J. Z. Gros, S. Dobrišek, and N. Pavešić, "Exploiting representation plurality for robust and efficient face recognition," in *Proc. 22nd Int. Electrotechnical Comput. Sci. Conf.*, 2013, pp. 121–124.
- [76] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The casia nir-vis 2.0 face database," in *Proc. Comput. Vis. Pattern Recog. Workshops*, 2013, pp. 348–353.
- [77] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
- [78] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 593–600.
- [79] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 13–26.
- [80] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.

- [81] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2088–2095.
- [82] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2160–2167.
- [83] X. Tan and B. Triggs, "Fusing Gabor and lbp feature sets for kernel-based face recognition," in *AMProc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, 2007, pp. 235–249.
- [84] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. 3rd Int. Conf. Image Signal Process.*, 2008, pp. 236–243.
- [85] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [86] T. I. Dhamecha, P. Sharma, R. Singh, and M. Vatsa, "On effectiveness of histogram of oriented gradient features for visible to near infrared face matching," in *Proc. 22nd Int. Conf. Pattern Recog.*, 2014, pp. 1788–1793.
- [87] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
- [88] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (SPA-E) for face recognition across poses," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1883–1890.
- [89] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.



**Jiwen Lu** received the BEng degree in mechanical engineering and the MEng degree in electrical engineering from the Xian University of Technology, Xian, China, in 2003 and 2006, respectively, and the PhD degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an associate professor in the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He serves as an

associate editor of the *Pattern Recognition*, *Pattern Recognition Letters*, *Neurocomputing* and the *IEEE Access*, an Elected Member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and an Elected Member of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society, respectively. He was a recipient of the National 1000 Young Talents Plan Program in 2015. He is a senior member of the IEEE.



**Venice Erin Liong** received the BS degree from the University of the Philippines Diliman, Quezon City, Philippines, in 2010, and the MS degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon City, South Korea, in 2013. She is working toward the PhD degree in the Interdisciplinary Graduate School, Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore. Her research interests include computer vision and pattern recognition. She is a student member of the IEEE.



**Jie Zhou** received the BS and MS degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. He is an associate editor of the *International Journal of Robotics and Automation* and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.

▷ For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/publications/dlib](http://www.computer.org/publications/dlib).